[BIOINFO 2025]

Metagenome-assembled genomes enhance bacterial read decontamination and variant calling in oral samples

Zunu An[1#], Jun Hyung Cha[1#], Kyu Ha Lee[2,3,4,*], and Insuk Lee[1,5,*]

[1]*Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University*
[2]*Department of Epidemiology, Harvard T.H. Chan School of Public Health*
[3]*Department of Nutrition, Harvard T.H. Chan School of Public Health*
[4]*Department of Biostatistics, Harvard T.H. Chan School of Public Health*
[5]*DECODE BIOME Co., Ltd.*
**Corresponding author: insuklee@yonsei.ac.kr (I.L.); klee@hsph.harvard.edu (K.H.L)*

Whole genome sequencing (WGS) offers advantages over DNA chip-based genotyping, typically using blood-derived DNA. However, saliva and buccal samples—popular in direct-to-consumer tests—suffer reduced accuracy because of oral bacterial contamination. Decontamination strategies using decoy bacterial genomes yielded limited improvements, likely because they cover only a subset of oral bacteria with available isolate genomes. To overcome this, we developed a decontamination pipeline leveraging metagenome-assembled genomes (MAGs). Concordance analysis of variant calling between blood and matched oral samples confirmed the superiority of MAG-augmented decontamination over conventional methods relying mainly on isolate genomes. Although the underlying mechanism remains unclear, it particularly improves variant calls in GC-rich regions, recovering many likely pathogenic variants. Additionally, we demonstrate that certain bacterial genomic regions mimic human regions with clinically relevant variants, potentially confounding genotyping. These results highlight the need for MAG-based bacterial read decontamination to achieve accurate personal genotyping from non-invasive, self-collected oral samples.